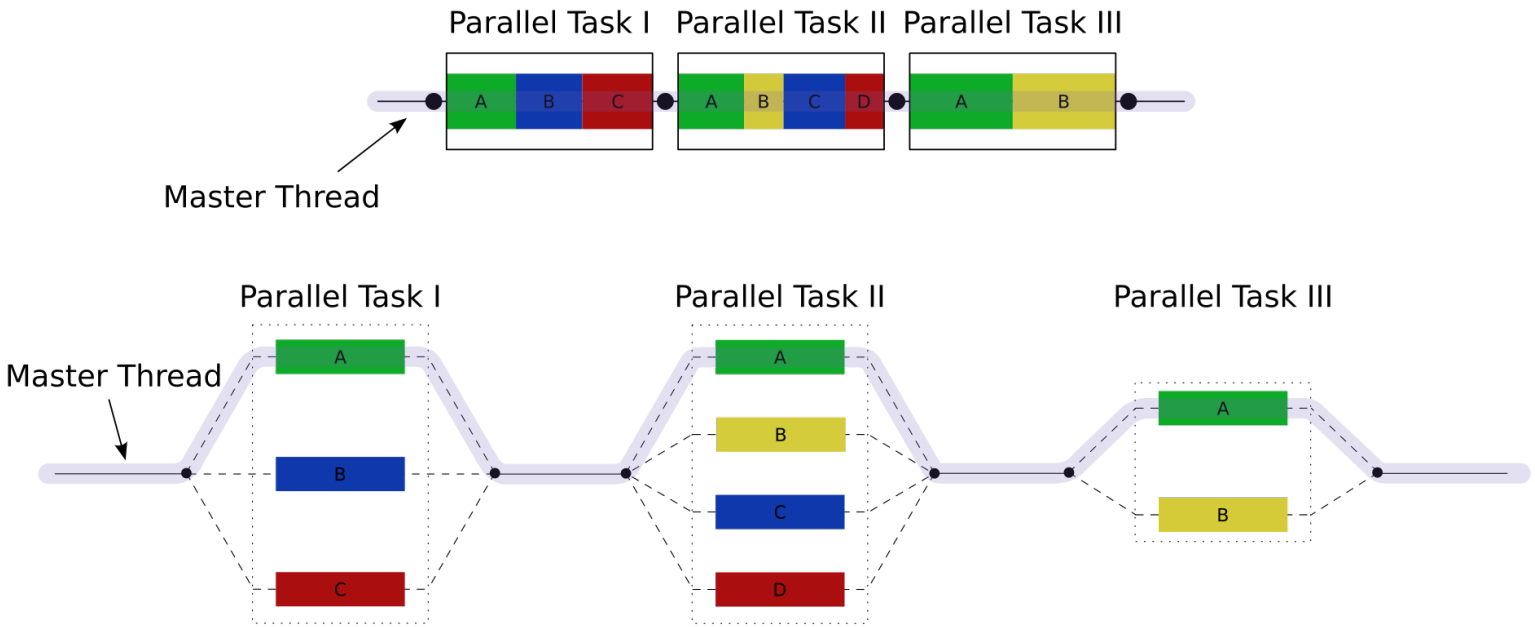
Parallel Programming with OpenMP

Introduction: OpenMP Programming Model

- Thread-based parallelism utilized on shared-memory platforms
- Parallelization is either explicit, where programmer has full control over parallelization or through using compiler directives, existing in the source code.
- Thread is a process of a code is being executed. A thread of execution is the smallest unit of processing.
- Multiple threads can exist within the same process and share resources such as memory

Introduction: OpenMP Programming Model

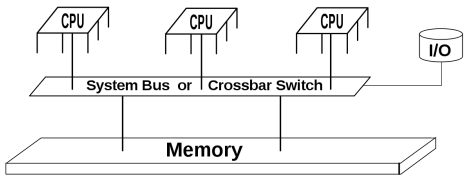
- Master thread is a single thread that runs sequentially; parallel execution occurs inside parallel regions and between two parallel regions, only the master thread executes the code. This is called the fork-join model:



Shared memory allows immediate access to all data from all processors without explicit communication.

Shared memory:

- multiple cpus are attached to the BUS
- all processors share the same primary memory
- the same memory address on different CPU's refer to the same memory location
- CPU-to-memory connection becomes a bottleneck: shared memory computers cannot scale very well



OpenMP versus MPI

OpenMP (Open Multi-Processing):

- easy to use; loop-level parallelism
- non-loop-level parallelism is more difficult
- limited to shared memory computers
- cannot handle very large problems

An alternative is MPI (Message Passing Interface):

- require low-level programming; more difficult programming
- scalable cost/size
- can handle very large problems

OpenMP core structure

```c
#include <omp.h>
#include <stdio.h>
#include <stdlib.h>

int main (int argc, char *argv[])
{
int nthreads, id;

/* Fork a team of threads giving them their own copies of variables.
   Make the values of nthreads and id private to each thread */

#pragma omp parallel private(nthreads, id)
  {

  /* Obtain thread number */
  id = omp_get_thread_num();
  printf("Hello World from thread = %d\n", id);

  /* Only master thread does this */
  if (id == 0)
    {
    nthreads = omp_get_num_threads();
    printf("Number of threads = %d\n", nthreads);
    }

  }  /* All threads join master thread and disband */
    return 0;
}
```

OpenMP core structure

```c
#include <omp.h>
#include <stdio.h>
#include <stdlib.h>

int main (int argc, char *argv[])
{
int nthreads, id;

/* Fork a team of threads giving them their own copies of variables.
   Make the values of nthreads and id private to each thread */

#pragma omp parallel private(nthreads, id)
  {

  /* Obtain thread number */
  id = omp_get_thread_num();
  printf("Hello World from thread = %d\n", id);

  /* Only master thread does this */
  if (id == 0)
    {
    nthreads = omp_get_num_threads();
    printf("Number of threads = %d\n", nthreads);
    }

  }  /* All threads join master thread and disband */
    return 0;
}
```

Beginning of the parallel region

Serial region

Fork a team of threads

Parallel region: all threads execute this

Compiling:

export OMP_NUM_THREADS=4

gcc -o hello_omp -fopenmp hello_omp.c

OpenMP C Directives

A directive has a name followed by clauses

```
#include <omp.h>

int omp_get_num_threads(void)
\* in bash, export OMP_NUM_THREADS=8 *\

#pragma omp parallel default(shared) private(beta,pi)

#pragma omp parallel for
#pragma omp critical
#pragma omp single
#pragma omp barrier


#pragma omp parallel for reduction(+:sum)


double omp_get_wtime(void);
```

OpenMP C Directives: Reduction

A private copy for each list variable is created and initialized for each thread. At the end of the reduction, the reduction variable is applied to all private copies of the shared variable, and the final result is written to the global shared variable.

Reduction

```
double avg, sum=0.0, A[MAX]; int i;
#pragma omp parallel for private ( sum )
        for (i = 0; i <= MAX ; i++)
                sum += A[i];

avg = sum/MAX;
```

Flow dependence

```
double avg, sum=0.0, A[MAX]; int i;
#pragma omp for reduction(+ : sum)
        for (i = 0; i <= MAX ; i++)
                sum += A[i];

avg = sum/MAX;
```

Dependence removed

```
#pragma omp critical
        sum=sum+x
```

Homework: Matrix multiplication

Review / Compile / Run the matrix multiplication code

```c
#include <omp.h>
#include <stdio.h>
#include <stdlib.h>

#define n 10                /* number of rows in matrix a */
#define k 10                /* number of columns in matrix a */
#define m 10                /* number of columns in matrix b */

int main (int argc, char *argv[])
{
int tid, nthreads, i, j, k;
double  a[n][k], b[k][k], c[n][k];

/* Begin a parallel region scoping all variables */
#pragma omp parallel shared(a,b,c,nthreads) private(tid,i,j,k)
  {
  tid = omp_get_thread_num();
  if (tid == 0)
    {
    nthreads = omp_get_num_threads();
    printf("Starting matrix multiple example with %d threads\n",nthreads);
    printf("Initializing matrices...\n");
    }
  /*** Initialize matrices ***/
  #pragma omp for
  for (i=0; i<n; i++)
    for (j=0; j<k; j++)
      a[i][j]= i+j;
  #pragma omp for
  for (i=0; i<k; i++)
    for (j=0; j<k; j++)
      b[i][j]= i*j;
  #pragma omp for
  for (i=0; i<n; i++)
    for (j=0; j<k; j++)
      c[i][j]= 0;

  /* Matrix multiply sharing iterations on outer loop */
  printf("Thread %d starting matrix multiply...\n",tid);
  #pragma omp for
  for (i=0; i<n; i++)
    {
    printf("Thread=%d did row=%d\n",tid,i);
    for(j=0; j<k; j++)
      for (k=0; k<k; k++)
        c[i][j] += a[i][k] * b[k][j];
    }
  }
  /* End of parallel region */

/*** Print results ***/
printf("******************************************************\n");
printf("Result Matrix:\n");
for (i=0; i<n; i++)
  {
  for (j=0; j<k; j++)
    printf("%6.2f   ", c[i][j]);
  printf("\n");
  }
printf("******************************************************\n");
return(0)

}
```

Performance with OpenMP

- The factor by which the time to solution can be improved compared to using only a single processor is called speedup.
- It is critical to parallelize the large majority of a program.
- Every time the program invokes a parallel region or loop, it incurs a certain overhead for going parallel.
- In addition to the costs of invoking the parallel loop and executing the barrier, cache and synchronization effects can greatly increase the cost.

Load balancing

- If some threads have to do more work than others, performance will suffer.
- Consider the problem of scaling the elements of a sparse matrix; if some rows have many more non-zero terms than others, load balancing can be a problem with static schedules.
- The load balancing problem can be solved by using a dynamic schedule.
- There are also costs to using dynamic schedules.
- A static schedule can usually achieve good load distribution in situations where the amount of work per iteration is uniform or if it varies in a predictable fashion

Scaling a dense, triangular matrix.

```
for (i = 0; i < n - 1; i++) {
    for (j = i + 1; j < n; j++) {
      a[i][j] = c * a[i][j];
    }
}
```

Scalability metrics

- Algorithmic limitations
- Serialized executions
- Startup overhead
- Communication

Scalability metrics

The overall size of the problem is $T^s = s + p$, where s is the serial (nonparallelizable) part and p is the perfectly parallelizable fraction.
Strong scaling: solving the same problem on N workers will require a runtime of

$$T^p = s + p/N$$

where N is the number of workers.
Weak scaling: scale the problem size

$$T^p = s + pN$$

Application speedup can be defined as the quotient of parallel and serial performance for fixed problem size.

Scalability metrics

Amdahls law, speedup for a parallel program using N processors, in which s is the non-parallelizable work, is given by:

$$P^p = \frac{1}{s + \frac{1-s}{N}}$$

Scalability: So-called strong scalability is in effect the same as speedup.
A code that shows perfect speedup presents strong scalability:

$$S^p = \frac{1}{s + \frac{1-s}{N}}$$

For a fixed problem size, scalability is limited.

Scalability metrics

In the case of weak scaling

$$P^p = S^p = \frac{s + (1 - s) * N^{\alpha}}{s + (1 - s) * N^{\alpha - 1}}$$

Weak scaling allows unlimited performance.
Parallel efficiency = speedup/N

$$\epsilon = s/N + 1 - s$$

Matrix multiplication

Homework: Matrix multiplication

- Analyze the speedup and efficiency of the parallelized code.
- Vary the size of your matrices from 250, 500, 750, 1000, and 2000 and measure the runtime with one thread.
- For each matrix size, change the number of threads from 2,4,8., and plot the speedup versus the number of threads. Compute the efficiency.
- Explain whether or not the scaling behavior is as expected.